Learning to Fuse Disparate Sentences

Micha Elsner and Deepak Santhanam

School of Informatics
University of Edinburgh

Department of Computer Science
Brown University

June 24, 2011

The problem: text-to-text generation

Input

The bodies showed signs of torture.

They were left on the side of a highway in Chilpancingo, in the southern state of Guerrero, state police said.

Output

The bodies of the men, which showed signs of torture, were left on the side of a highway in Chilpancingo, state police told Reuters.

Motivation

Humans fuse sentences:

- Multidocument summaries (Banko+Vanderwende '04)
- Single document summaries (Jing+McKeown '99)
- **Editing (this study)**

Motivation

Humans fuse sentences:

- Multidocument summaries (Banko+Vanderwende ‘04)
- Single document summaries (Jing+McKeown ‘99)
- **Editing (this study)**

Previous work: multidocument case:

- Similar sentences (*themes*)
- Goal: summarize common information

(Barzilay+McKeown ‘05), (Krahmer+Marsi ‘05), (Filippova+Strube ‘08)...

Our task setting

Sentences fused by professional editors–
Related by discourse, but...

Content is not usually similar!

Our task setting

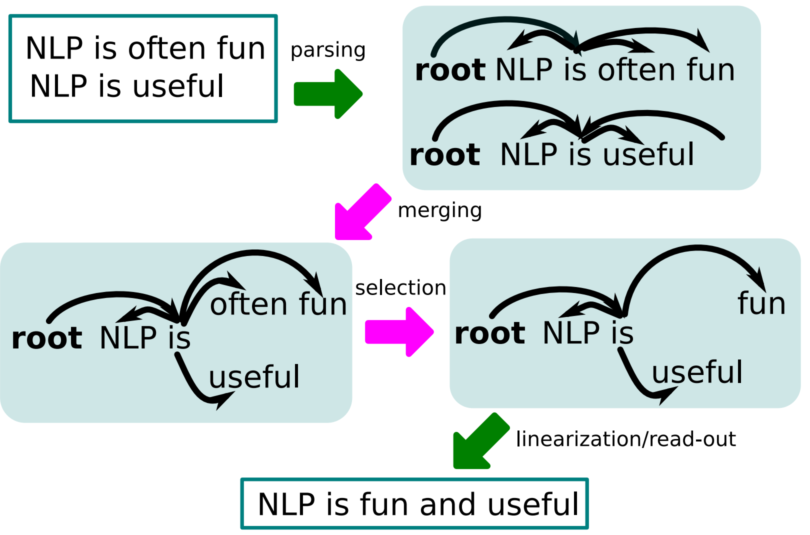
Sentences fused by professional editors–
Related by discourse, but...

Content is not usually similar!

Editing data:

- Naturally occurring dataset
- Probably more similar to single-document summary
- Poses problems for standard approaches

Generic framework for sentence fusion



Issues with the generic framework

Selection

What content do we keep?

- Convey the editor's desired information
- Remain grammatical

Merging

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

Learning

Can we learn to imitate human performance?

Issues with the generic framework

Selection

What content do we keep?

- Convey the editor's desired information
  - **Requires discourse; not going to address**
- Remain grammatical

Merging

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

Learning

Can we learn to imitate human performance?

Issues with the generic framework

Selection

What content do we keep?

- Convey the editor's desired information
  - **Requires discourse; not going to address**
- Remain grammatical
  - **Constraint satisfaction** (Filippova+Strube '08)

Merging

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

Learning

Can we learn to imitate human performance?

Issues with the generic framework

Selection

What content do we keep?

- Convey the editor's desired information
  - **Requires discourse; not going to address**
- Remain grammatical
  - **Constraint satisfaction** (Filippova+Strube '08)

Merging

Which nodes in the graph match?

Dissimilar sentences: correspondences are noisy!

Contribution: Solve jointly with selection

Learning

Can we learn to imitate human performance?

Issues with the generic framework

Selection

What content do we keep?

- Convey the editor's desired information
  - **Requires discourse; not going to address**
- Remain grammatical
  - **Constraint satisfaction** (Filippova+Strube '08)

Merging

Which nodes in the graph match?

Dissimilar sentences: correspondences are noisy!

Contribution: Solve jointly with selection

Learning

Can we learn to imitate human performance?

Contribution: Use structured learning

Overview

- Motivation
- Setting up the problem
- Fusion as optimization
  - Jointly finding correspondences
  - Staying grammatical
- Learning to fuse
  - Defining an objective
  - Structured learning
- Evaluation

Overview

Motivation

Setting up the problem

Fusion as optimization
- Jointly finding correspondences
- Staying grammatical

Learning to fuse
- Defining an objective
- Structured learning

Evaluation

The data

500 article pairs processed by professional editors:

Novel dataset courtesy of Thomson Reuters

THOMSON REUTERS

Each article in two versions: **original** and **edited**

We align originals with edited versions to find:

- 175 split sentences
- 132 merged sentences
- ... treat both as fusion examples

The content selection problem

Which content to select?

(Daume+Marcu '04), (Krahmer+al '08)

Fake content selection

Use simple dynamic programming to align input with truth...
Provide true alignments to both **system** and **human judges**.

Input

Uribe appeared unstoppable after the rescue of Betancourt.

His popularity shot to over 90 percent, but since then news has been bad.

True output

Uribe appeared unstoppable and his popularity shot to over 90 percent.

Overview

Motivation

Setting up the problem

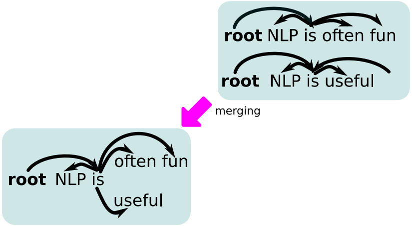
Fusion as optimization
- Jointly finding correspondences
- Staying grammatical

Learning to fuse
- Defining an objective
- Structured learning

Evaluation

Merging dependency graphs

Previous:



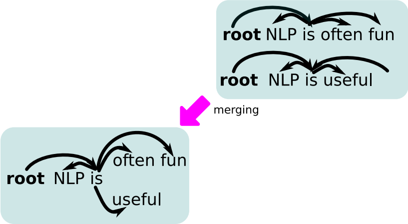
Merge nodes deterministically:

- Lexical similarity
- Local syntax tree similarity

For disparate sentences,
these features are noisy!

Merging dependency graphs

Previous:

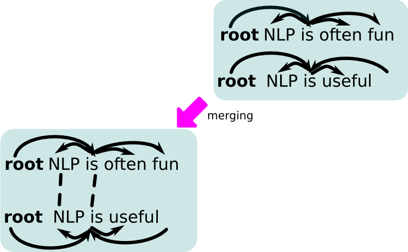


Merge nodes deterministically:

- Lexical similarity
- Local syntax tree similarity

For disparate sentences, these features are noisy!

Our work:

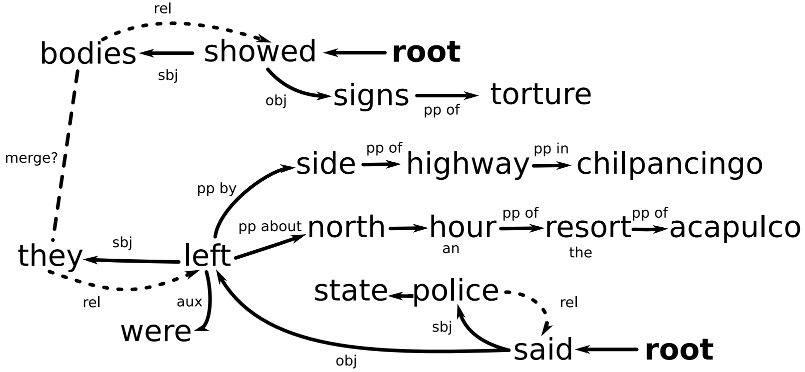


Soft merging: add **merge arcs** to graph

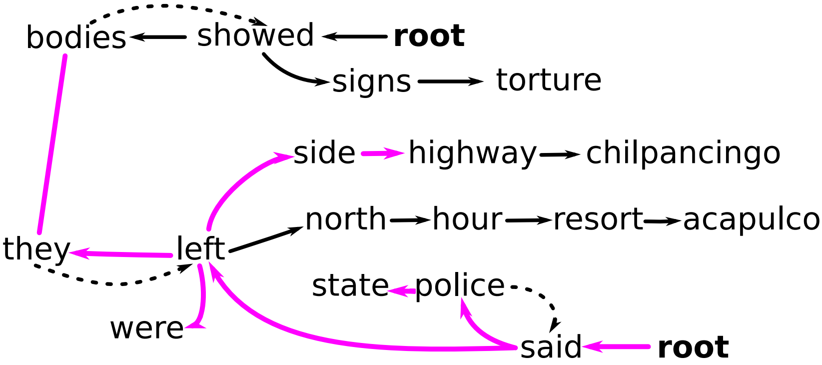
System decides whether to use or not!

Add relative clause arcs between subjects and verbs

(Alternates “police said” / “police, who said”)



A fused tree: a set of arcs to keep/exclude

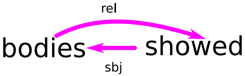


"State police said the bodies were left by the side of a highway"

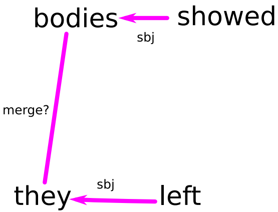
Constraints

Not every set of selected arcs is valid...

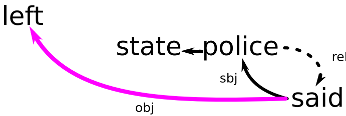
Unconnected fragment



Cycle



Merged node with two heads



Missing argument (subject)

Solving with ILP

Integer Linear Programming (ILP)

Generic method for hard combinatorial problems
Well-studied practical solutions (Ilog CPLEX)

Our ILP based on (Filippova+Strube '08), generalized for soft merging...
Similar setup for sentence compression (Clarke+Lapata '08)

Very efficient for this size problem

Overview

Motivation

Setting up the problem

Fusion as optimization

- Jointly finding correspondences
- Staying grammatical

Learning to fuse

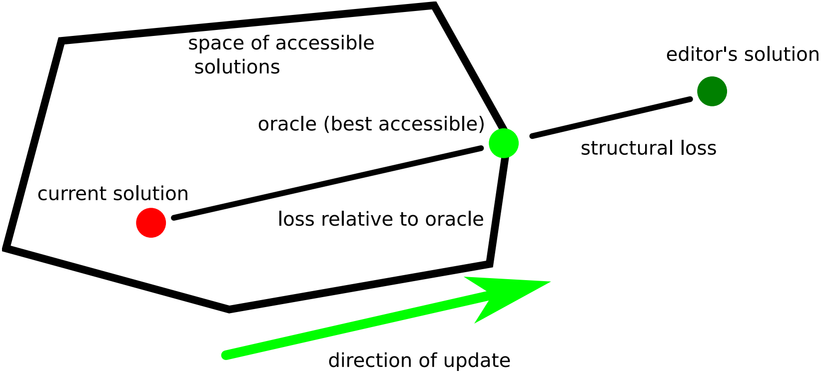
- Defining an objective
- Structured learning

Evaluation

How to fuse?

ILP tells us what fusions are allowed...
  Weights in objective tell us which ones are good.

Recipe for structured learning, (Collins '02),others:

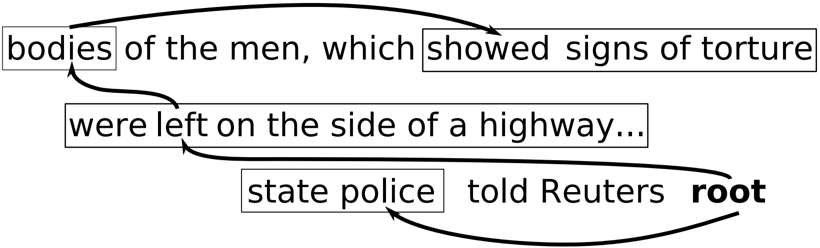


Loss function

Measure similarity to the editor's sentence...

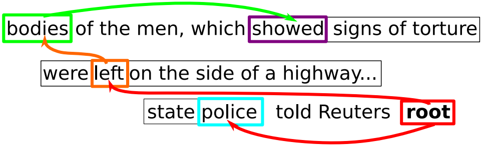
- Not just lexically (the editor can paraphrase, we can't!)

Look at **connections between** the retained content

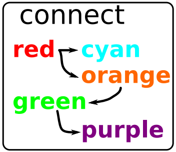


Finding the oracle

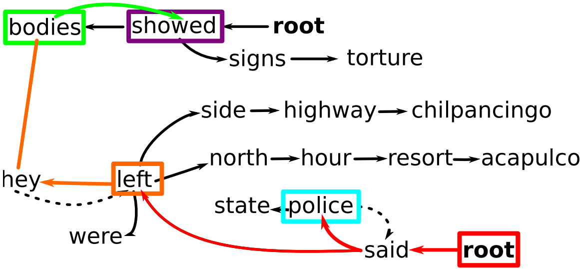
Match this structure:



editor's sentence



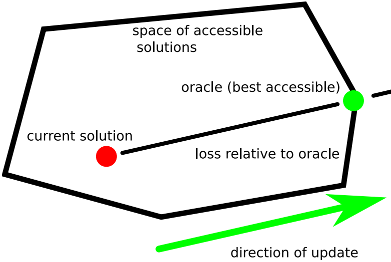
On this graph:



our graph

Optimizing

- Loss function/oracle:
  - Just shown; connections between regions
- Features:
  - POS tags, dependency labels, lexical resources
- Update rule:
  - Averaged perceptron



Overview

Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

Human

The editor's fused sentence

Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

Human

The editor's fused sentence

Readability upper bound

Our parsing and linearization on the editor's sentence

Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

Human

The editor's fused sentence

Readability upper bound

Our parsing and linearization on the editor's sentence

"and"-splice

All input sentences, spliced with the word "and"

Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

Human

The editor's fused sentence

Readability upper bound

Our parsing and linearization on the editor's sentence

"and"-splice

All input sentences, spliced with the word "and"

System

Our system output
Only abstractive system we tested

Readability

System	**Avg**
Editor	4.6
Readability UB	4.0
“And”-splice	3.7
System	3.1

- Poor linearization: gap of .6
- System: additional loss of .9
- Average system score still 3, “fair”

Content

System	Avg
Editor	4.6
Readability UB	4.3
“And”-splice	3.8
System	3.8

- Score close to 4, “good”

Comparison with “and”-splice

“and”-splice content scores comparable to ours, but...

- Spliced sentences too long
  - 49 words vs human 34, system 33
- Our system has more extreme scores

	1	2	3	4	5	Total
“And”-splice	**3**	43	60	57	**103**	266
System	**24**	24	39	58	**115**	260

Good

Input

The bodies showed signs of torture.

They **were left on the side of a highway in Chilpancingo**, in the southern state of Guerrero, **state police** said.

Our output

The bodies who showed signs of torture were left on the side of a highway in Chilpancingo state police said.

Good

Input

The suit **claims** the company **helped fly terrorism suspects abroad to secret prisons.**

Holder's **review was disclosed the same day as Justice Department lawyers repeated a Bush administration** state-secret **claim in a lawsuit against a Boeing Co unit.**

Our output

Review was disclosed the same day as Justice Department lawyers repeated a Bush administration claim in a lawsuit against a Boeing Co unit that helped fly terrorism suspects abroad to secret prisons.

Not very good

Our system

Biden a veteran Democratic senator from Delaware that Vice president-elect and Joe had contacted to lobby was quoted by the Huffington Post as saying Obama had made a mistake by not consulting Feinstein on the Panetta choice.

Better parsing/linearization

Vice President-elect Joe Biden, a veteran Democratic senator from Delaware who had contacted...

Not very good

Our system

The White House that took when Israel invaded Lebanon in 2006 showed no signs of preparing to call for restraint by Israel and the stance echoed of the position.

Missing arguments

took, *position*

Conclusion

- Naturally occurring data
- Find correspondences jointly with selection
- Supervised structured learning

Future work

New data:

- Classic similar-sentence fusion (McKeown '10 corpus)
- Single-document summary

Better techniques:

- Paraphrasing rules

Code available

bitbucket.org/melsner/sentencefusion

Acknowledgements

Thompson-Reuters: Alan Elsner, Howard Goller, Thomas Kim
BLLIP labmates: Eugene Charniak, Stu Black, Rebecca Mason, Ben Swanson
Funds: Google Fellowship for NLP
All of you!